Audio Coding Using Variable-Depth Multistage Quantization *

Faouzi Kossentini[1], Michael Macon[2†], and Mark J. T. Smith [2]

[1]Department of Electrical Engineering,
University of British Columbia,
Vancouver BC V6T 1Z4, Canada

[2]School of Electrical and Computer Engineering,
Georgia Institute of Technology,
Atlanta, GA 30332-0250, USA

† M. Macon is currently with the Oregon Graduate Institute, Portland, OR, 97291-1000

Abstract

In this paper, an algorithm for high quality coding of 48 kHz sampled audio signals is presented. The signal is first converted to a frequency-domain representation using the Modulated Lapped Transform. Perceptual irrelevancies in the signal are removed by employing frequency-domain human auditory masking properties to derive a masked quantization noise threshold. The noise threshold is then smoothed using simple morphological operations to account for limitations in the frequency resolution of the transform. This smoothed threshold is then used as the input to a multistage uniform scalar quantization scheme that shapes quantization noise to fit below the masking threshold. Higher order entropy coding is applied to the output of the quantizer, simultaneously exploiting dependencies in the time, frequency, and stage dimensions. A variable number of quantization stages is used for each transform coefficient. However, the nature of the formulation is such that no side information is required to send the number of stages. Through an informal experiment, the audio quality of the new coder was found to be better than that of the MPEG layer I coder and roughly equivalent to that of the MPEG layer II coder.

*This work was supported by the Natural Sciences and Engineering Research Council of Canada under grant # OGP-0187668 and NASA.

1 Introduction

Digital compression and coding of high fidelity audio signals has become a key technology in the development of cost-effective multimedia systems [1, 2, 3, 4, 5, 6]. Most audio coding algorithms rely on (a) removal of statistical redundancies in the signal and (b) exploitation of masking properties of the human auditory system to "hide" distortions in the coded signal. Transform and subband coders provide a convenient framework for time-frequency domain signal analysis and coding based on these two principles. The most notable examples are the coders that are compatible with the MPEG standard audio coding layers I-III [6]. Such coders are based on subband/transform decomposition, perceptual masking, scalar quantization, and simple block companding (layers I and II) or multi-dimensional Huffman coding (layer III).

In the coding algorithm proposed here, we employ variable-depth multistage quantization together with a perceptual transform to exploit noise masking redundancies [7]. The front-end perceptual transform system is based on the Modulated Lapped Transform (MLT) [8] as well as all the well-known auditory modeling features. The unique feature of this work is the utilization of variable-depth multistage uniform quantizers in conjunction with high order entropy coding. The construction avoids completely the need to send side information updates about the number of stages (or depth level) in the multistage quantizer.

2 Auditory Masking Model

Masking is a psychoacoustic phenomenon that renders low level signals concentrated in a given frequency region inaudible in the presence of higher level signals at neighboring frequencies. The masking model used here is similar to the well-known method developed by Johnston [3]. It relies on the computation of (a) summations of signal energy over frequency regions corresponding to *critical bands* of the auditory system [9], (b) a cochlear *spreading function* which describes the effects of signal energy in one critical band on masking in adjacent bands, and (c) a measure of *tonality* of the signal. From this frequency domain analysis, a *masking threshold* is obtained and used to provide a perceptual upper bound on the level of quantization noise that can be tolerated at each frequency for a given short-time signal segment. Since the MLT is here applied to relatively long audio frames (e.g, 1024 samples used to compute 512 coefficients per frame), it provides a fine frequency resolution. Thus, the critical band energies are computed directly

from the MLT subband outputs in each frame.

At low frequencies, the critical bandwidth is quite narrow. In this frequency region, the sidelobe amplitudes and main lobe width of the transformed MLT basis functions are significant in comparison to the threshold shape, since leakage from adjacent frequencies can violate the masking threshold, as shown in Figure 1. Thus, the masking curve must be constrained to take the basis function shape into account. To overcome this problem, a morphological *opening* [10] of the masking threshold is performed, using the magnitude squared of the transformed basis function as a structuring element. This operation results in a masking threshold that is less than or equal to the original threshold in magnitude, but obeys the frequency resolution constraints of the MLT.

3 Quantization and Entropy Coding

By applying an N-point MLT ($N = 512$ coefficients computed from 1024 input samples) and the masking model to M (e.g., 4, 8, 16) consecutive audio frames, $M \times N$ significant[1] coefficients and their associated masking thresholds are obtained. The $M \times N$ matrix of coefficients represents the short-time frequency characteristics of the audio signal. As illustrated in Figure 2, the $M \times N$ matrix is divided into subbands. The subbands are then quantized independently using a multistage residual uniform quantizer and entropy coded by an arithmetic encoder [11]. The arithmetic encoder employs probabilities that are obtained adaptively using a finite state machine (FSM), as described below.

3.1 Quantization

Each stage quantizer consists of *three* levels. The only information about the multistage quantizer required by the decoder is the maximum value x_{max}, the minimum value x_{min}, and the average value x_{avg} of the coefficients in each subband. Both the encoder and the decoder can then construct the first stage quantizer by choosing the reconstruction levels to be $y_0 = x_{avg}$, $y_1 = x_{min} + \frac{1}{3}(x_{avg} - x_{min})$, and $y_2 = x_{max} - \frac{1}{3}(x_{max} - x_{avg})$. The maximum magnitude of the error allowed by the first stage is $\Delta_1 = \frac{1}{3}\max(x_{avg} - x_{min}, x_{max} - x_{avg})$. The ith ($i = 2, 3, \ldots$) subsequent stage quantizer is reconstructed by setting its levels to $y_0 = 0$, $y_1 = -\frac{2}{3}\Delta_i$, and

[1] Ignoring the coefficients above $M = 464$ (frequencies > 21 kHz) does not affect the perceptual quality of the reproduced audio signal.

$y_2 = \frac{2}{3}\Delta_i$, where Δ_i is obtained using Δ_1 and the recursion equation $\Delta_i = \frac{\Delta_{i-1}}{3}$. Since the average value of the residual data is usually very close to zero, setting the middle level to zero introduces no significant additional distortion. However, this has the advantage that quantization can be stopped any time the distortion falls below the associated masking threshold value, and zeros are then output by any subsequent stage quantizers. By encoding the sequence of zeros, the encoder indicates that adequate precision has been obtained without explicitly sending side information to the decoder.

Following the above construction, the value of the first coefficient of a particular subband is assigned one of the three levels of the first stage quantizer by way of two comparisons. If the distortion, which is the positive difference between the coefficient value and the value of its assigned level, is below the masking threshold, quantization of the difference is not required. Otherwise, the difference is quantized using a new stage quantizer that is constructed as described above. This process continues until the distortion falls below the masking threshold. The other subband coefficients are quantized using current multistage quantizers, assuming zeros as outputs for unnecessary stage quantizers and constructing new quantizers when required. Clearly, the number of stage quantizers can vary from one subband to another, and can be determined once the last subband coefficient is quantized.

3.2 Entropy Coding

As illustrated in Figure 3, the multistage uniform quantizer outputs as many multiresolution approximations of the subband coefficients as the number of stage quantizers. Such approximations appear as slices of a spectrogram that is shaped by the masking noise. Each slice consists of three grey levels representing the outputs of the corresponding stage quantizer for the subband coefficients. Statistical dependencies exist in the time, frequency, and stage dimensions, and these can be efficiently and effectively exploited by using a simple algorithm that is based on the statistical modeling method described in [12, 13, 14].

The modeling algorithm used in this work is based on a FSM whose state transitions are determined from some previously decoded stage quantizer outputs or symbols. The number and coordinates of the best few $(4-6)$ conditioning symbols change from one subband to another, but they vary slowly as a function of the characteristics of the audio signal. Thus, such information is

determined off-line using a fast implementation of the tree-structured searching method described in detail in [13, 14]. More specifically, we first select a sufficiently large region of support that consists of many neighboring (in time, frequency, and stage) conditioning symbols. Using a large training set representing different types of audio material, we then compute approximate values of conditional entropies for all the neighboring conditioning symbols. Only those symbols with conditional entropies significantly smaller than the first order entropy are selected. Such an approximation is suboptimal. However, it is shown experimentally to perform very well, and it requires only approximately three adds/shifts per subband coefficient.

The encoding of the stage symbols consists of (1) identifying the closest state and associated table of probabilities and (2) arithmetically encoding according to the probabilities. That is, a vector of values of the conditioning symbols is mapped to a vector in a state book. The state book usually contains a few vectors, each representing a state. The state points to a table of frequencies used by the arithmetic coder to identify the appropriate code space for the current symbol.

To explain the mapping procedure, let $\mathbf{x} = (x_1, \ldots, x_n)$ be the vector of values of the n selected conditioning symbols. Similarly, let $\mathbf{y}^s = (y_1^s, \ldots, y_n^s)$ be a state book vector associated with a state s. The vector $\mathbf{x}$ is compared to each state vector $\mathbf{y}^s$, and if the two vectors are equal, the associated state s is selected. If the vector $\mathbf{x}$ is different from all current state vectors, it is then added to the state book. As the state book size reaches a predetermined maximum value, the least popular state vector is deleted and the new vector is added. For each state vector, the distribution of the table probabilities can be adapted to the local statistics of the audio signal. In this work, we follow the simple strategy described in [11], where a table frequency corresponding to the current coded symbol is incremented, and the frequencies in the particular table are halved once their sum exceeds a predetermined maximum value.

4 Experimental Results

In an informal subjective comparison test, the quality of our coder was compared with that of the MPEG layers I-II coders whose implementation can be found at *http://drogo.cselt.stet.it/mpeg*. The intent of this test was to gain a rough benchmark of the audio quality in a working coder. Better results may be obtained by optimization of the front-end transform and perceptual model.

We presented each of five test signals SUZANNE VEGA (f1), TRACY CHAPMANN (f2), CHIMES (f3), FIREWORKS (f4), and SAX-TRUMPET (f5) (each 10–15 sec duration) to 25 volunteer subjects via headphones. The signals were sampled at 48 kHz and monophonic, but presented to both ears. The test subjects were graduate students with varied experience in subjective audio assessment. In each test case, two signals were presented in a random order: (*i*) the test signal coded by the MPEG layer I or II coder and (*ii*) the same signal coded using our coder. Each listener was then asked to choose the signal he/she preferred, after listening to the pair as many times as desired. The playback level was adjusted by each subject once, only at the beginning of each session.

Our audio coder employs the MLT to transform audio frames of length 1024, (overlapped by 50%) to generate 512 coefficients. Among the 512 MLT coefficients per frame, only the $N = 464$ lowest in frequency are quantized and coded. By grouping $M = 32$ audio frames, a 32×464 matrix of coefficients is obtained. The matrix is divided into 4 rectangular subbands of length 32 in the time dimension and widths 58, 58, 116, and 232 in the frequency dimension.

Although the number of stage quantizers is determined during the quantization process, it still cannot exceed a practical maximum value of 10. The maximum number of conditioning symbols is set to 10 and the maximum allowed number of states is set to 100. However, our implementation employs at most 4 symbols and 16 states, while still achieving essentially the same level of compression performance. The table probabilities are approximated with one-byte frequency counts, where the lowest probability is $\frac{1}{256}$ and the highest probability is $\frac{255}{256}$.

Table 1 summarizes the results of subjective comparison of our coder and the MPEG layer I coder. As is clear from the tables, a majority of listeners preferred our coder over MPEG layer I. It also should be noted that that our coder performs better at 80 kilobits per second (kbps) than the MPEG layer I coder at 96 kbps. However, Table 2 reveals that the subjective quality of our coder is roughly equivalent to that of the MPEG layer II at 64 and 80 kbps. As shown in the tables, the test signal (f3), which consists of a set of chimes being struck, was more difficult for our coder than some of the other signals, and some unmasked quantization noise was audible. This indicates that further tuning of the psychoacoustic model may be necessary.

Our coder is comparable in terms of computations to the MPEG layers I-II. First, the MLT filter bank requires approximately 26 multiplies per sample. Second, our quantization involves only an average of about 3 compare and subtract operations. Third, our current FSM imple-

mentation requires only a few compare operations. Finally, even adaptive arithmetic coding can now be performed very efficiently, using many of the fast implementations developed recently [CITE ?].

5 Conclusions

A transform coding scheme based on the Modulated Lapped Transform, a perceptual noise masking model, and a new quantization/modeling algorithm has been presented. The perceptual model is able to take into account the limitations of the transform resolution in setting a noise detection threshold. From this threshold, a multistage uniform quantizer is used to shape quantization noise to fit the masking curve. The FSM model used to encode the output of the quantizer exploits statistical redundancies in the time, frequency, and stage dimensions, resulting in effective compression of the audio signal.

The proposed audio coder is roughly equivalent in quality with MPEG layer II and performs better than MPEG layer I at the tested bit rates. Its computation demands are small, and it is amenable to simple hardware and software implementations.

References

[1] J. Princen, A. Johnson, and A. Bradley, "Subband/transform coding using filterbank designs based on time-domain aliasing cancellation," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*, pp. 2161–2164, IEEE, April 1987.

[2] R. N. J. Veldhuis, M. Breeuwer, and R. V. D. Waal, "Subband conding of digital audio signals without loss of quality," *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*, pp. 2209–2012, May 1989.

[3] J. D. Johnston, "Transform coding of audio signals using perceptual noise criteria," *IEEE Journal on Selected Areas in Communications*, vol. 6, pp. 314–323, February 1988.

[4] W.-Y. Chan and A. Gersho, "High fidelity audio transform coding with vector quantization," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*, (Albuquerque, New Mexico, USA), pp. 1109–1112, Apr. 1990.

[5] M. R. Soleymani, "New tandem source-channel trellis coding scheme," *IEEE Transactions on Speech and Audio Processing*, vol. 2, pp. 24–28, Jan. 1994.

[6] ISO/IEC 13818-2—ITU-T Rec. H.262, *Generic Coding of Moving Pictures and Associated Audio Information: Video.* ISO/IEC, 1995.

[7] F. Kossentini, M. Macon, and M. J. T. Smith, "Audio coding using variable-depth multistage quantizers," in *Data Compression Conference*, (Snowbird, UT, USA), Mar. 1996.

[8] H. S. Malvar, *Signal Processing with Lapped Transforms.* 685 Canton Street, Norwood, MA 02062: Artech House, Inc., 1992.

[9] E. Zwicker and H. Fastl, *Psychoacoustics: Facts and Models*, ch. 4: Masking. Springer-Verlag, Berlin, 1990.

[10] P. Maragos and R. W. Schafer, "Morphological systems for multidimensional signal processing," *Proc. of the IEEE*, vol. 78, pp. 690–710, April 1990.

[11] I. H. Witten, R. M. Neal, and J. G. Cleary, "Arithmetic coding for data compression," *Communications of the ACM*, vol. 30, pp. 520–540, 1987.

[12] F. Kossentini, W. Chung, and M. Smith, "Subband image coding with jointly optimized quantizers," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*, vol. 4, (Detroit, MI, USA), pp. 2221–2224, May 1995.

[13] F. Kossentini, W. Chung, and M. Smith, "Conditional entropy-constrained residual VQ with application to image coding," *Transactions on Image Processing: Special Issue on Vector Quantization*, vol. 5, pp. 311–320, Feb. 1996.

[14] F. Kossentini, W. Chung, and M. Smith, "A jointly optimized subband coder," *IEEE Trans. on Image Processing*, vol. 5, pp. 1311–1323, Sept. 1996.

List of Tables

List of Figures

Coder at 64 kbps	f1	f2	f3	f4	f5
MPEG layer I (%)	4	4	56	20	12
Our Coder (%)	96	96	44	80	88

Coder	f1	f2	f3	f4	f5
MPEG layer I at 96 kbps (%)	16	20	60	44	44
Our Coder at 80 kbps (%)	84	80	40	56	56

Coder	64 kbps	96 kbps
MPEG layer I (%)	19	36
Our Coder (%)	81	64

Table 1: Subjective quality comparisons between our coder and MPEG layer I. Numbers are percentages of the listeners who selected the coder given in the same row.

Coder at 64 kbps	f1	f2	f3	f4	f5
MPEG layer II (%)	80	68	56	48	24
Our Coder (%)	20	32	44	52	76

Coder at 80 kbps	f1	f2	f3	f4	f5
MPEG layer II (%)	72	40	72	44	52
Our Coder (%)	28	60	28	56	48

Coder	64 kbps	80 kbps
MPEG layer II (%)	55	56
Our Coder (%)	45	44

Table 2: Subjective quality comparisons between our coder and MPEG layer II. Numbers are percentages of the listeners who selected the coder given in the same row.

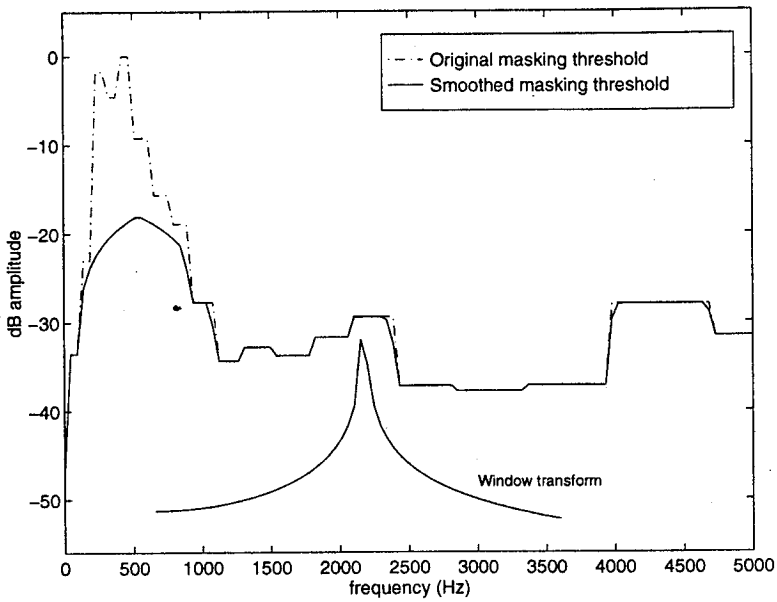

Figure 1: Morphological smoothing of masking threshold. The analysis window transform magnitude squared (transform of a single MLT basis function) is used as the structuring element for the smoothing operation, which accounts for limited frequency resolution in the transform.

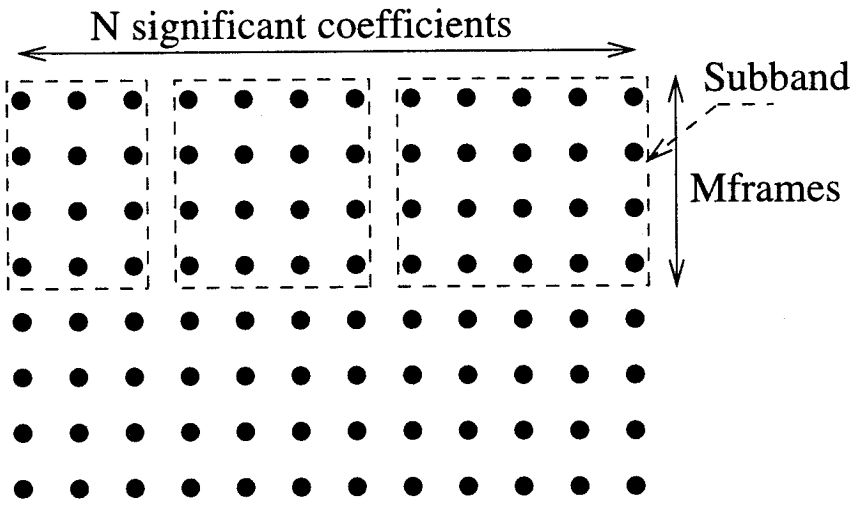

Figure 2: MLT coefficient grouping procedure

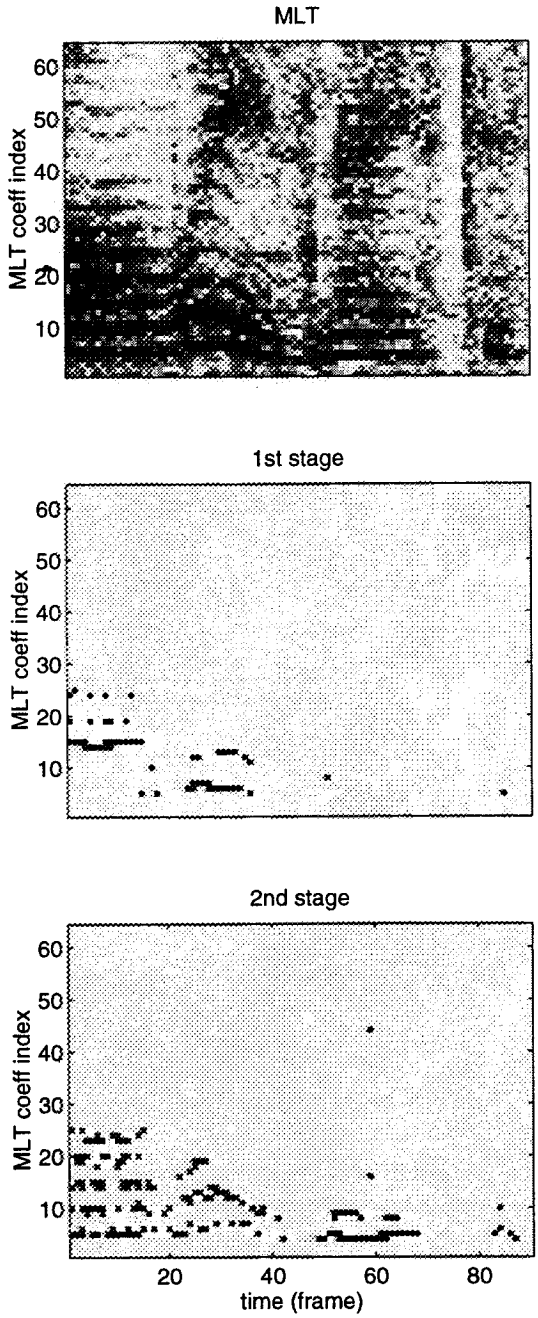

Figure 3: Top: portion of MLT "spectrogram" of audio signal. Middle and bottom: quantizer stage outputs – white, grey, and black represent stage quantizer output levels. Note the dependencies in the time, frequency, and stage dimensions.